GO-ESSP 2011 Workshop:

Parallel Analysis of GeOscience Data Status and Future

Jeff Daily

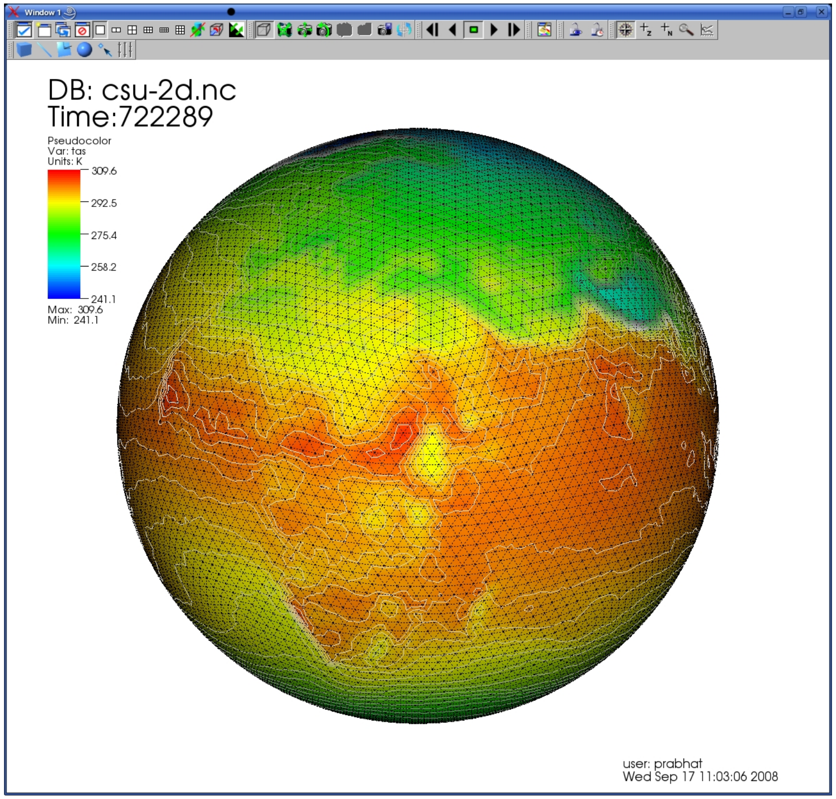
PI: Karen Schuchardt
in collaboration with Dave Randall et al

http://svn.pnl.gov/gcrm/wiki/pagoda
pagoda-dev@googlegroups.com

Pacific Northwest
NATIONAL LABORATORY

Motivation

- Data sets approaching PB range
  - GCRM = 1.4 PB
    - 4 km resolution
    - 3 hourly, 1 simulated year
  - CCSM4 = 100 TB
    - 0.1 degree ocean
    - 0.24 degree atmosphere
    - 1 simulated century
  - 1 PB / 10GB/s = 28 hours
  - 1 PB / 300MB/s > 40 days
  - IO bandwidth is the bottleneck
  - 64bit offsets needed to describe file
- Support GCRM's geodesic grid
  - Semi-structured
  - Explicit topology variables



DB: csu-2d.nc
Time:722289

Pseudocolor
Var: tas
Units: K

Max: 309.6
Min: 241.1

user: prabhat
Wed Sep 17 11:03:06 2008

Image courtesy of Prabhat at LLNL using VisIt tool on GCRM data.

Approach

- Parallel IO part of the design
  - Parallel NetCDF
  - NetCDF4/HDF5
- Data-parallel versus task parallel
  - Data distribution and communication using Global Arrays library
  - Task parallel is still possible
- C++ API for custom analysis
  - Similar to Java NetCDF API
  - IO and Array operations
- Command-line tools for immediate use
  - Mimic an established interface
  - Drop-in replacement for NetCDF Operators
- Support geodesic and regular grids

Original Data Model

- CF compliant cell-based data model
  - Needed variables defined on corners, edges
  - Redundancy in the data (cell bounds are logically shared)
  - Cannot traverse cells for visualization (e.g. isolines) or analysis

```
dimensions:
    cell = 10240
    nv = 6
variables:
    float center_lon(cell)
    float center_lat(cell)
    float corner_lon(cell, nv)
    float corner_lat(cell, nv)
```

- Enhancements to NetCDF Operators
  - Auxiliary coordinate support for cell-based grids
  - Performance improvements for hyperslabbing (subsetting)
  - No support for explicit topology

```
ncks -X lon_min,lon_max,lat_min,lat_max
          vs.
ncks -d lon,min,max -d lat,min,max
```

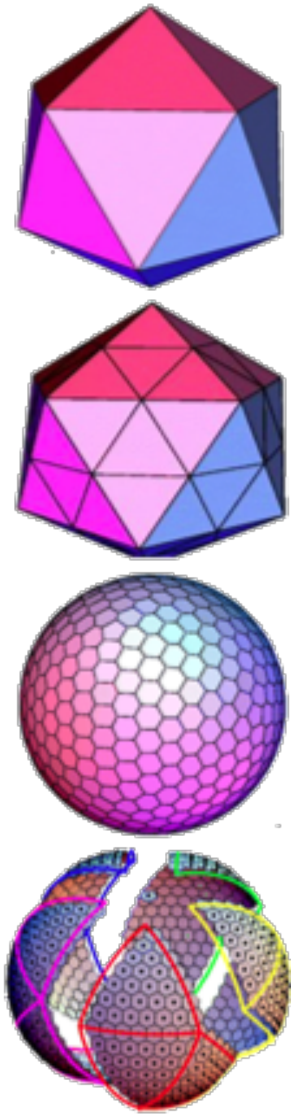
Data Model

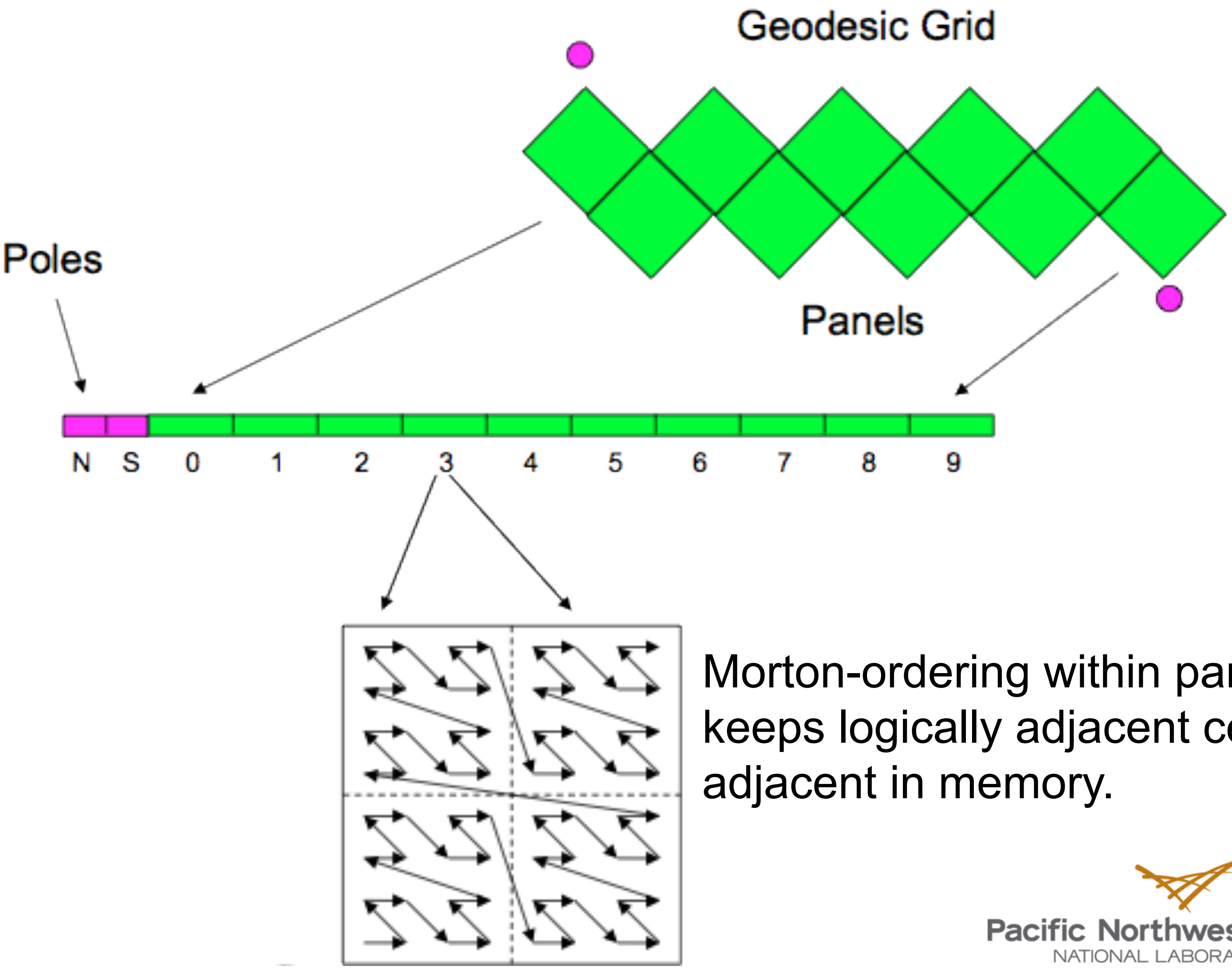
```
// Dummy scalar for grid discovery
int grid;
    grid:standard_name = "grid”;
    grid:external_ref = "some uri”;
    grid:cell_type = "hex”;
    grid:index_start = 0s;
    // topology references
    grid:cell_edges   = "cell_edges”;
    grid:cell_corners = "cell_corners";
    grid:cell_cells   = "cell_neighbors”;
    grid:edge_corners = "edge_corners";
    // geometry references
    grid:coordinates_cells   = “center_lon center_lat”;
    grid:coordinates_corners = “corner_lon corner_lat”;
    grid:coordinates_edges   = “edge_lon edge_lat";
```

Data Model (cont.)

```
dimensions:
  cells      =  41943042
  corners    =  83886080
  edges      = 125829120
  time       = UNLIMITED
  layers     = 24
  interfaces = 25
variables:
  float temperature(time,cells,layers)
  float wind(time,edges,interfaces)
  float center_lon(cells)
  float center_lat(cells)
  float corner_lon(corners)
  float corner_lat(corners)
  float edge_lon(edges)
  float edge_lat(edges)
  int   cell_corners(cells, cellcorners=6)
  int   cell_edges(cells,celledges=6)
  int   edge_corners(edges,edgecorners=6)
  int   cell_neighbors(cells,cellneighbors=6)
```

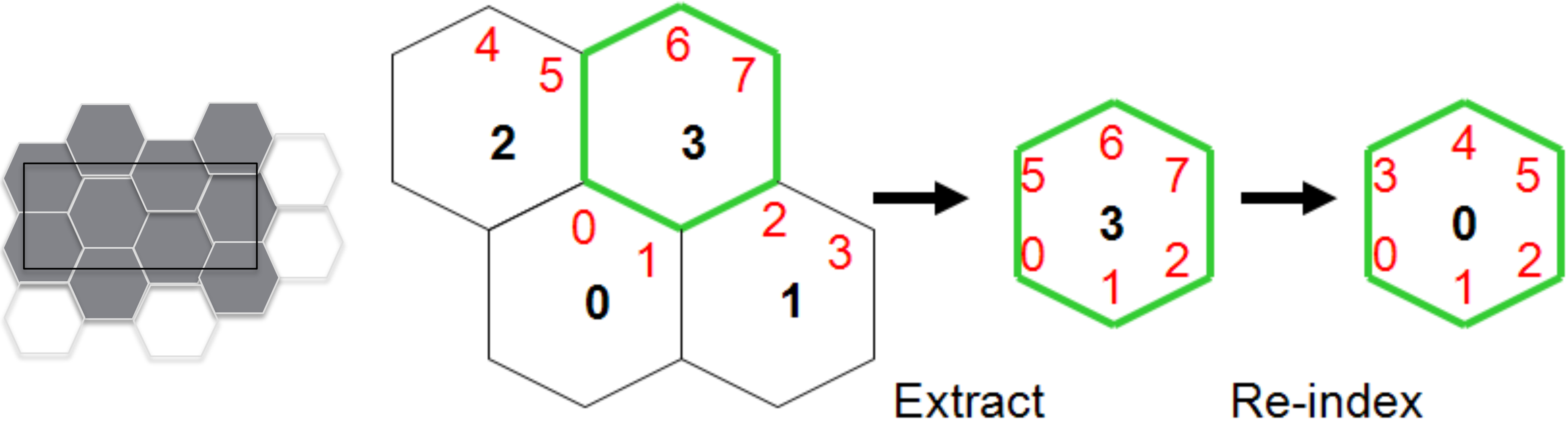
Geodesic Grid





Morton-ordering within panels keeps logically adjacent cells adjacent in memory.

Subsetting the Geodesic Grid



- Can’t use start+count or strided NetCDF API (data is unstructured)
- Mask-based (arbitrary subset regions)
- Maintain whole cells (renumber topology variables)
- “subsetter” was the first pagoda command-line tool

NetCDF Operators

NCO	What it does
ncks	subsetting, text display
ncra	record average
ncea	ensemble average
ncwa	weighted average
ncbo	binary arithmetic
ncflint	file interpolation
ncrcat	record concatenation
ncecat	ensemble concatenation
ncrename	rename vars/dims
ncatted	edit attributes
ncpdq	permute dimensions
ncap/ncap2	scripted processor

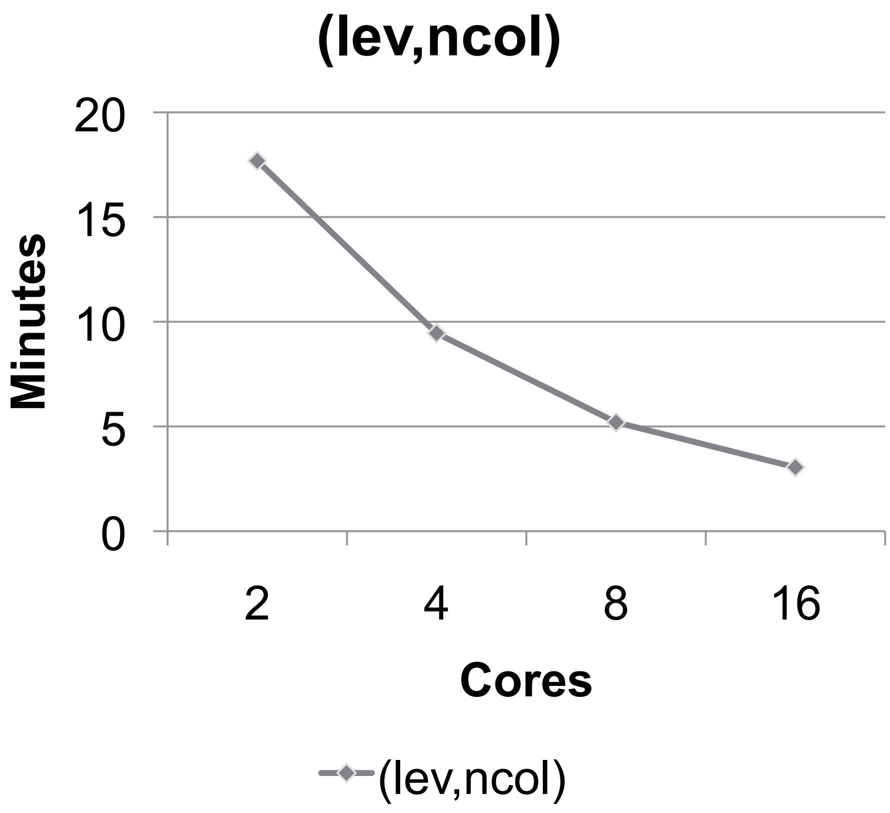
- Serial command-line tools
  - Use interactively
  - As part of a script
- Task-parallel versions
  - No parallel IO
- Extremely portable
- Not intended as library
  - "To my knowledge, though, only NCO programs use libnco"
  - Installs libnco but no headers
  - Potential for code reuse

Pagoda Command-line Tools

NCO	Pagoda
ncks	pgsub
ncra	pgra
ncea	pgea
ncwa	(soon, v0.7)
ncbo	pgbo
ncflint	pgflint
ncrcat	N/A*
ncecat	N/A*
ncrename	
ncatted	
ncpdq	
ncap/ncap2	

- Current version is 0.6
- Output verified against NCO
  - Tested GCRM data
    - 8km resolution
  - Tested against ANL data
    - 1/8 degree CAM HOMME
    - 19 8.5GB files (15 variables each)
    - 19 2.5GB files (4 variables each)
  - Assumes NCO infallible
- Scriptable (but not as simple)
- *Don’t concatenate, aggregate

pgra Strong Scaling on eureka.alcf.anl.gov

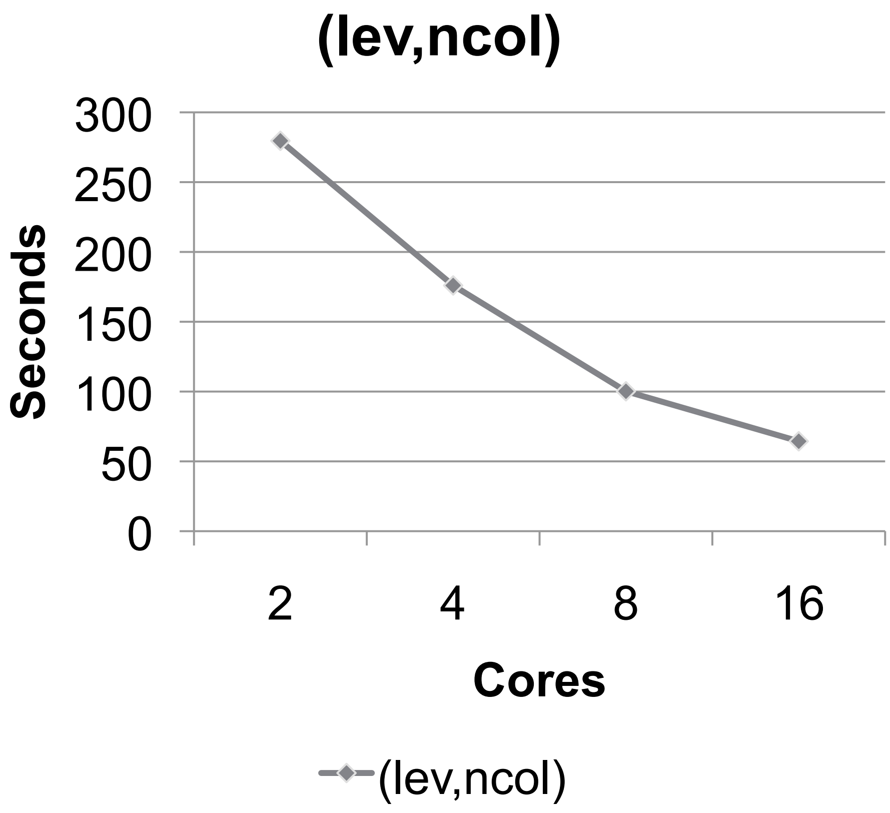


- 19 8.5GB files
  - 15 variables each
  - 1/8 degree CAM HOMME
  - (lev=26,ncol=3110402)
- Timings

```
 1-core ncra 16:06
 2-core pgra 17:41
 4-core pgra  9:27
 8-core pgra  5:12
16-core pgra  3:03
```

Thanks to Sherri Mickelson at ANL for the data and for performing these runs.

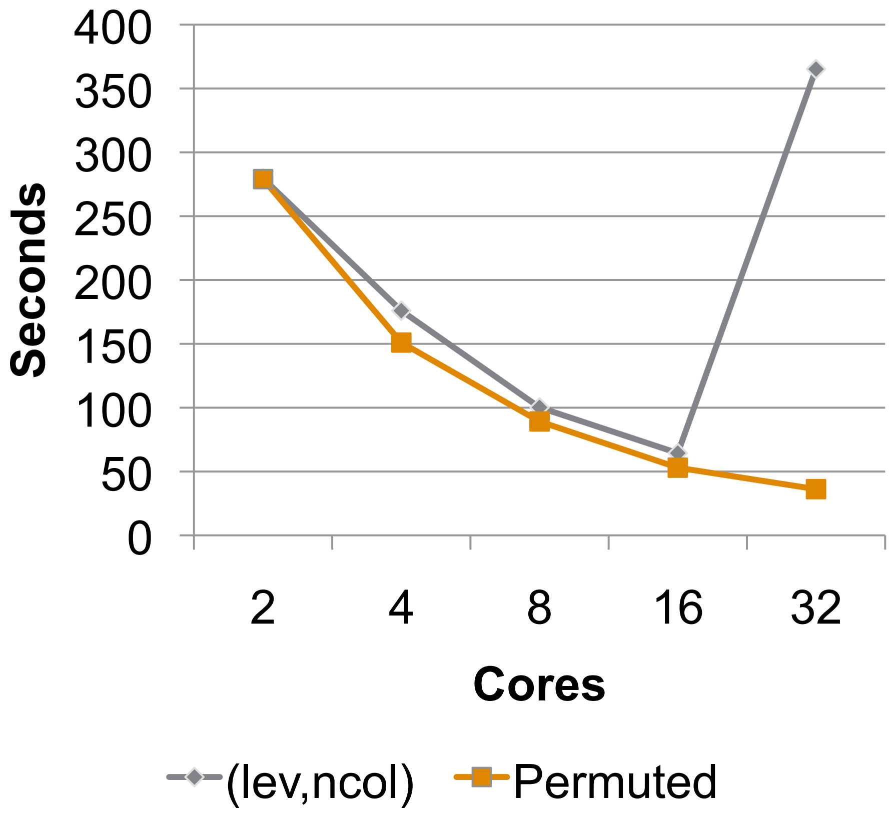
pgra Strong Scaling on eureka.alcf.anl.gov



- 19 2.5GB files
  - 4 variables
  - 1/8 degree CAM HOMME
  - (lev=26,ncol=3110402)
- Timings
  - `1-core ncra 4:54`
  - `2-core pgra 4:39`
  - `4-core pgra 2:56`
  - `8-core pgra 1:40`
  - `16-core pgra 1:04`

Thanks to Sherri Mickelson at ANL for the data and for performing these runs.

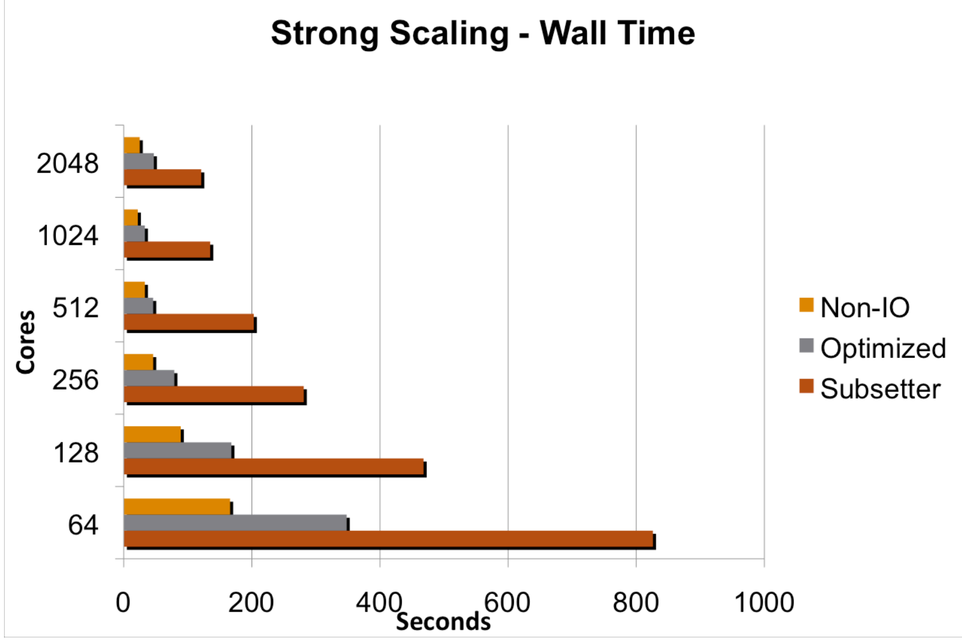
pgra Strong Scaling on eureka.alcf.anl.gov

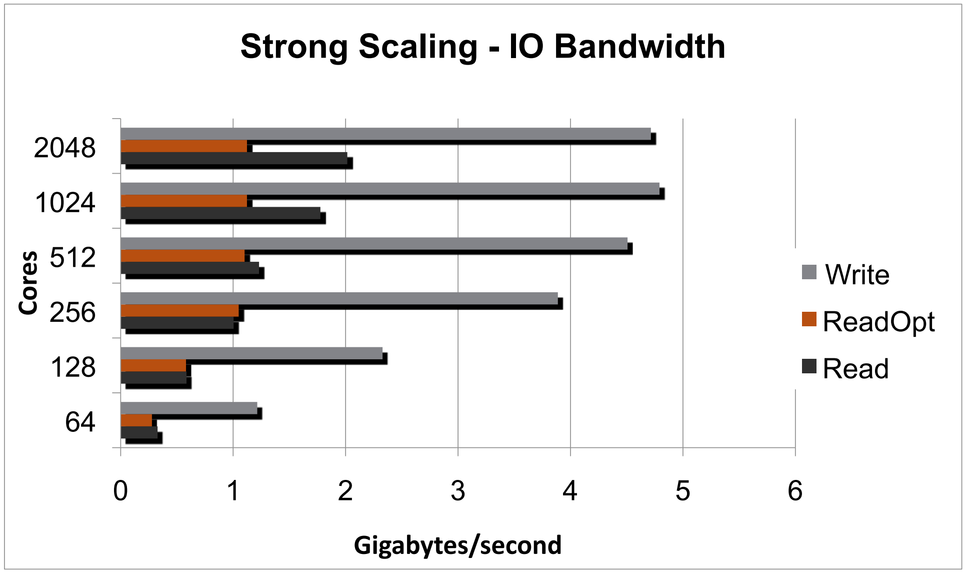


- Scalability depends on dimension order and data distribution
  - Using pnetcdf-1.2.0
  - netcdf4/hdf5 may not be impacted
  - (lev=26,ncol=3110402)
  - Permuted i.e. (ncol,lev)
  - Only distributing first dimension when smaller than number of cores

Thanks to Sherri Mickelson at ANL for the data and for performing these runs.

pgsub Strong Scaling on franklin.nersc.gov



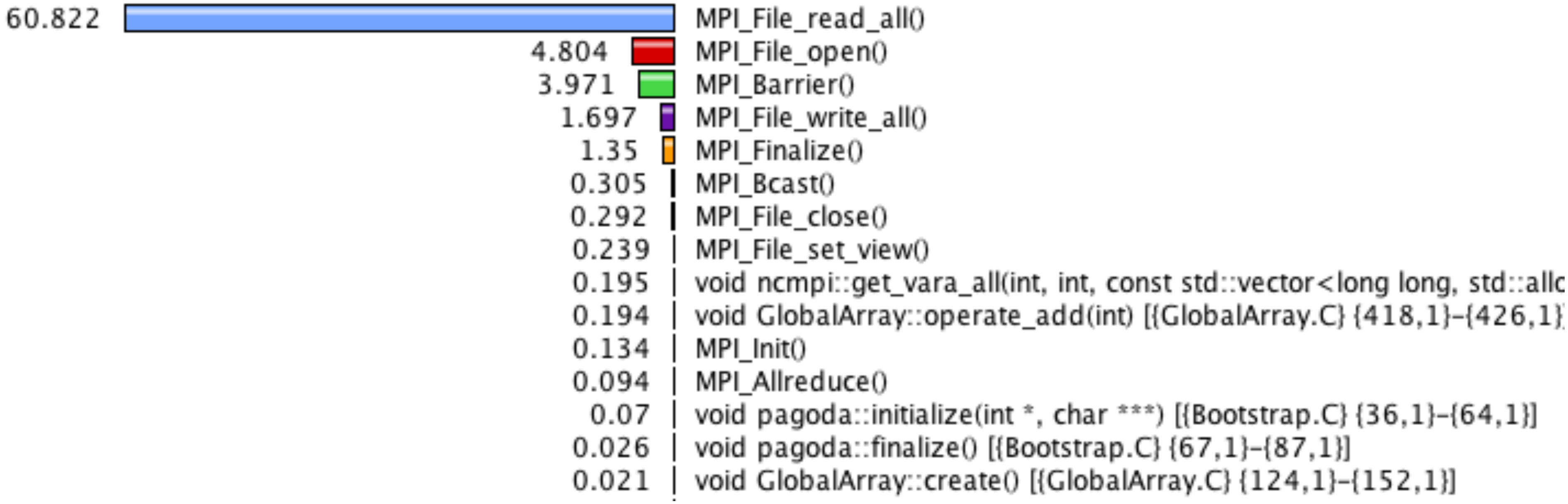


- Shown to scale up to 2K cores
- Shows that IO is a major bottleneck
- Write bandwidth nearly 5GB/s on franklin
- Our first optimization shows importance of efficient use of IO

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

pgra Flat Profile via TAU



60.822	MPI_File_read_all()
4.804	MPI_File_open()
3.971	MPI_Barrier()
1.697	MPI_File_write_all()
1.35	MPI_Finalize()
0.305	MPI_Bcast()
0.292	MPI_File_close()
0.239	MPI_File_set_view()
0.195	void ncmpi::get_vara_all(int, int, const std::vector<long long, std::allc
0.194	void GlobalArray::operate_add(int) [{GlobalArray.C} {418,1}-{426,1}
0.134	MPI_Init()
0.094	MPI_Allreduce()
0.07	void pagoda::initialize(int *, char ***) [{Bootstrap.C} {36,1}-{64,1}]
0.026	void pagoda::finalize() [{Bootstrap.C} {67,1}-{87,1}]
0.021	void GlobalArray::create() [{GlobalArray.C} {124,1}-{152,1}]

- Using franklin.nersc.gov, 1K cores, 40 OSTs
- Variable: wind(time=8*17, edges=7864302, layers=26)
  - ~0.76 GB per timestep
  - >100 GB total (not including grid variables)
- Overwhelming majority of time spent in IO
  - Using non-blocking pnetcdf API to aggregate small IO
  - Working with IO library developers to optimize

Future

- “make it easy” – A higher level API
- New language bindings? Python (via Cython)? Fortran?
- Handle more conventions e.g. scale_value, add_offset
- Finish pgwa (ncwa)
- Grid interpolation, possibly integrating ESMF code
- Hide IO latency by mixing IO and computation
- Other operators e.g. pdq, ncap/ncap2
  - What if header isn’t big enough and data is too large?
  - What if pnetcdf’s “CDF5” format is used?
- We need more users and user input on what’s needed
  - Already in use/testing by CSU, ANL, NCAR
  - In testing as replacement for NCO tools in nightly NCAR script

Acknowledgements

This research is supported by the U. S. Department of Energy's Office of Science under the Scientific Discovery through Advanced Computing (SciDAC) program.

SciDAC
Scientific Discovery through Advanced Computing

U.S. DEPARTMENT OF ENERGY

Office of Science
U.S. DEPARTMENT OF ENERGY

Thanks

GO-ESSP 2011 Workshop:

Parallel Analysis of GeOscience Data Status and Future

Jeff Daily

PI: Karen Schuchardt
in collaboration with Dave Randall et al

http://svn.pnl.gov/gcrm/wiki/pagoda

pagoda-dev@googlegroups.com

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*